圖書與資訊學刊
55(Nov '05)33-50
ISSN 1023-2125

索引典之自動化建置與視覺化

Automatic Construction and Visualization of a Thesaurus

林 頌 堅
Sung-Chien Lin
世新大學資訊傳播學系助理教授
Assistant Professor, Department of Information and Communication Studies
Shih-Hsin University
E-mail：scl@cc.shu.edu.tw

【摘要 Abstract】

本論文描述自動化索引典建置與資訊視覺化的方法與結果。我們提出一個自動化方法，依據論文中的文字資訊，從無到有地建置索引典。這個方法利用統計訊息從論文的文字資訊中選取具有代表性的術語，偵測術語之間的概念關係，並利用資訊視覺化技術，將索引典資訊表示成直覺且資訊豐富的圖形。本論文以政大圖書與資訊學刊為對象，進行索引典建置與資訊視覺化的試驗，並且根據不同的用途，提出各種檢索與瀏覽的使用方式。試驗結果說明了這個方法的可行性與效果。

This paper describes automatic methods for thesaurus construction and information visualization and their testing results. A method is developed to construct a thesaurus from scratch, using only the textual materials in the examined domain of papers as the main information resource. All the terms relevant to the domain are selected and their mutual conceptual relationships are detected, based on the statistical processing of the input texts. The selected terms and their conceptual relationships are applied to construct a thesaurus dedicated to the examined domain. Another method for visualizing information in the constructed thesaurus is also proposed by generating a set of graphs, in which the mapped positions of related terms are displayed in juxtaposition. These graphs are useful in showing the knowledge structure of the examined domain and very suitable for the applications of retrieval and browsing. In the study, on the basis of our proposed methods, we have performed tests on the textual materials selected from papers published by the periodical *Bulletin of Library and Information Science,* NCCU. Various usages of the graphs in the constructed thesaurus have been experimented in terms of different sorts of applications. The final results confirm the feasibility as well as the effectiveness of the two methods.

關鍵詞 Keyword

索引典建置　資訊視覺化　術語選取　術語關係偵測

Thesaurus construction；Information visualization；Term selection；Term relation detection

壹、緒論

索引典(Thesaurus)是資訊組織的重要資源之一，用來儲存某一特定領域的詞彙以及術語與術語之間的各種概念關係。比方說《ASIS Thesaurus of Information Science and Librarianship》是圖書資訊學領域的索引典，這個索引典蒐集了圖書資訊學相關的術語，並將這些術語依據各種概念關係加以組織，如 BT(Broader term)、NT(Narrower term)、RT(Related term)、UF(Use for)等等。這些關係中，BT 與 NT 構成了術語間「廣泛—特定」的階層式概念關係。若是一個術語是另一個術語的 BT，表示前者是後者的廣泛概念，後者則是前者的一種特定概念，並且以前者的 NT 來表示。比方說，“university libraries”是一種特殊的“libraries”，前者是後者的一種特定概念；因此，術語“libraries”是術語“university libraries”的 BT，而“university libraries”則是“libraries”的 NT。此外，兩個術語概念相關但不是「廣泛—特定」的關係，則可以用 RT 來表示它們相互間的關係(Rowley, 1992, p.255-256)。利用索引典中蘊藏的詞彙訊息，對文件資料進行索引，可以提高檢索的效能，使結果更加符合使用者的需求。比方說，在建立文件資料庫時，編目人員(indexers)可以根據索引典的詞彙和概念關係，選擇文件主題相關的術語來對文件進行索引；檢索者也可以依據索引典，利用符合需求的術語作爲問句進行查詢(Soergel, 1985, p.222)；因爲此時編目人員和檢索者雙方使用詞彙資訊的一致，可以提高檢索的準確率(Precision rate)。另外，當檢索獲得的資料太少時，可能是檢索的主題太過於特定，資料庫內的相關資料不多。此時使用者便可以從索引典的概念關係中，選取相關術語的 BT 與 RT 重新檢索，以提高檢索的回收率(Recall rate)。而且索引典中的術語代表了相關領域的重要概念，術語之間的概念關係便是領域的知識組織情形。因此，利用索引典可以提供使用者瀏覽與探索領域知識結構的全貌與細節，是初學者相當重要的參考資源。

既然索引典是資訊檢索與知識領域瀏覽等資訊組織應用上相當重要的資源，便需要有效率且系統化的製作方法。傳統上索引典多以人工方式建置，建立索引典的工作者需要閱讀大量的文獻，並且與相關領域的專家進行大量的訪談，從這些知識來源中取得各種足以代表領域重要概念的術語，並分析術語之間的概念關係(Rowley, 1992, p.269-270)。因此，索引典的建置需要付出極大的成本。當領域的發展十分迅速的時候，將有許多新的概念與相關術語不斷出現，索引典需要經常修改與維護，這種情形下更加需要龐大的專家知識與各種成本，此時明顯地可以看出人工建置方法的限制與問題。因此，電腦科學家與資訊科學家便提出多種自動化方法，嘗試利用電腦的快速處理能力與極大的記憶容量，協助建置索引典。由於近來科技研究領域的快速變動，加以電子論文及資料庫的急遽增加，提供了發展索引典自動化建置方法的資源與需求，這項技術便成爲資訊檢索研究與發展的重要方向。目前所提出的這些方法多利用術語相互間的詞彙訊息和語法關係來取得建構索引典所需的資訊或是利用術語在文件中出現的統計訊息之相似程度做爲概念關係的判定，而這些方法在某些特定的應用上也有極成功的效果，但仍然有許多值得進一步研究之處。

在索引典的資訊呈現方面，以紙本呈現索引典的內容，除了採用以字母爲排列方式的循序方式之外，多以主題爲排列的概念階層方式爲主。比方說，《ASIS Thesaurus of Information Science and Librarianship》便提供了這兩種呈現方式。近年來，資訊科技的進步產生了大量的電子文件，

許多索引典也以電子形式儲存並透過電腦螢幕呈現，目前常見的索引典電腦介面有階層選單(Hierarchical menu)模式(Sanderson & Croft, 1999)和網路模式(Tseng, 2002)。以階層選單模式來說，利用索引典的階層式樹狀結構，第一層的選單中列出表整個領域最上層概念的術語，做爲選項；在選擇每一個選項後，將會打開第二層的選單，選單中的選項是與第一層術語相關但意義較爲特殊的術語。以此類推，將索引典中所有的術語依據上下層的概念關係，放置入階層選單模式中(Sanderson & Croft, 1999)。使用者利用這個機制時，首先瀏覽代表最上層概念的術語，接著可以選擇與需求相關的一個術語展開，以便瀏覽與該術語相關但意義較爲特殊的術語，如此反覆展開，使用者可以了解術語之間廣泛或特殊的概念關係，並且每次僅注意於局部的資訊，避免使用者的認知能力發生過度負荷的情形。

但是目前的索引典呈現方式在詞彙資訊的取得上並不方便。比方說，紙本的索引典不易於瀏覽整個領域的知識結構，即便利用階層選單模式的電子索引典也非常困難。再者，當使用者需要比較兩個術語之間的概念關係時，除了直接上下與相關的概念關係之外，紙本索引典需要經過多次的翻閱，而電子索引典也需點選多次。因此目前的索引典呈現方式，不管是傳統的紙本或是新發展的電腦介面，不但資訊的取用方式不夠直覺，而且將對使用者的認知、理解和記憶等能力造成極大的負荷。上述的問題可以藉由資訊視覺化(Information visualization)的處理獲得解決。資訊視覺化是近年來電腦科學技術研發的重要方法(Card, Mackinlay & Shneiderman, 1999)，利用電腦強大的運算與繪圖能力，使得複雜而難以理解的大量資料，根據資料的特徵形成圖形，方便使用者解讀。若是可以透過資訊視覺化技術的處理，利用圖形介面來呈現索引典中蘊含的資訊，將領域中重要的術語分布表現在圖形上，所能呈現的資訊更爲豐富，並且在認知上更爲直覺而容易理解。使用者將可以一覽整個領域的知識組織，認識重要的研究主題。並提供放大(Zoom in)的功能，使得使用者可以觀看圖形上局部地區的詳細術語分布情形，了解術語和術語之間的關係，索引典的使用將可以更加方便與有效。

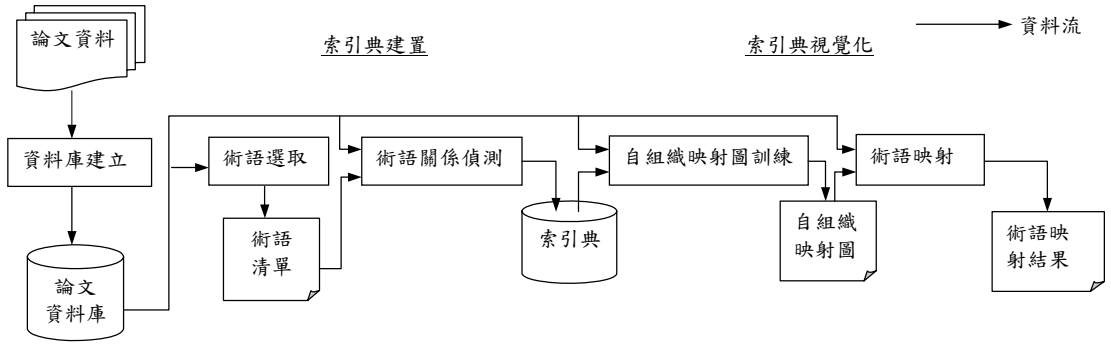
因此，本論文將進行索引典的自動化建置與資訊視覺化方法的研究，並且特別針對於發展迅速但缺乏索引典等資源的學術領域。在本研究中，領域是指某一學術社群(Scholarly community)成員所研究的知識範疇與結構，比方說圖書資訊學或性別研究等等。學術社群的成員從領域的知識範疇與結構中學習，了解社群所關注的研究問題，並且熟悉領域所認可的理論、研究方法與技術，而能夠進行相關的研究。換言之，領域的內容包含相關學術研究社群所關心的問題、進行研究所使用的理論、方法、技術和結果等等知識，以及各種知識之間的關連。當學術社群的成員在研究領域內相關問題時，所得到的結果與知識將經由論文加以陳述，並且透過相關的期刊或研討會發表。因此，作者在論文中記載了問題、理論、方法、技術等研究相關的資訊。在通過編輯與同儕審查(Peer review)的檢驗並在社群中進行傳播等論文發表的過程後，這些論文中所記載的研究資訊，比方說新問題、新方法和新技術等等，可以轉化爲領域中新的知識範疇與結構，而可以爲社群成員再加以利用。具有影響力的論文在被社群成員閱讀、認可而使用之後，作者用來指稱問題和方法等研究上特定概念的某些術語，將爲社群成員熟悉而在研究資訊的交流過程中使用。所以論文中的術語及其意義與領域知識的範疇與結構息息相關，透過論文的文字資訊分析將可以揭露領域知識的範疇與結構，作爲領域特定的索引典建置所需的資訊。

基於上述的想法，本研究中將提出以論文資料爲資訊來源的自動化索引典建置與索引典資訊視覺化方法，並爲了驗證上述方法的效果與可行性，將以政治大學圖書館所出版的《圖書與資訊學刊》裡的論文資料做爲研究對象，實際從論文的題名和摘要等文字資訊中抽取關鍵術語，自動建置索引典，並進行資訊視覺化。

本論文其餘部分的組織如下：第貳節中將說明索引典自動化建置與資訊視覺化的整體方法流程。第參節爲自動索引典建置之相關研究分析，作爲本研究中選擇索引典建置方法之依據。第肆節首先進行資訊視覺化方法的文獻探討，並說明本研究應用自組織映射圖技術呈現索引典內容的理由與方法。第伍節說明利用《圖書與資訊學刊》作爲研究對象，應用於本研究所提出的方法應用在《圖書與資訊學刊》論文的實驗，包括研究對象的範圍以及實驗所得到的結果等。最後，第陸節則是本文的結論。

貳、研究方法流程

考慮到許多學術研究領域缺乏現有的索引典，並且不希望借助過多的人力與成本，本研究使用的資源主要來自於領域相關的論文資訊，整個處理流程包含「索引典建置」和「索引典視覺化」兩個程序。索引典建置利用統計導向的方法，以領域相關的論文作爲資訊來源，選取具有代表性的術語，並推斷術語之間的概念關係。資訊視覺化處理則採用自組織映射圖(Self-organizing maps, SOM)技術(Kohonen, 1989)，將所有的術語映射到二維圖形上，產生出代表領域知識結構的圖形。圖一是本研究之整體方法的流程。



圖一：索引典視覺化的流程圖

在索引典建置和索引典視覺化等程序之前，首先需要蒐集領域相關的論文，建立論文資料庫。在論文資料庫儲存的資料，包括各論文的題名、摘要、甚至包含本文等文字資料。

索引典建置程序包括「術語選取」(Term selection)和「術語關係偵測」(Term relationship detection)兩個部分。術語選取對於論文文字資料進行統計，選取論文中以詞(Words)或詞組(Phrases)爲單元的中英文術語，並且這些術語必須同時具有主題代表性(Kageura & Umino, 1996)和鑑別性(Discrimination)(Crouch & Yang, 1992)等特性。術語關係偵測則是對於選取出來的術語利用它們出現的論文相對頻率推算術語之間的涵攝情形(Subsumption)，建立術語之間的上下層概念關係(Sanderson & Croft, 1999)。

接著進行索引典資訊視覺化，如前所述，本研究選擇自組織映射圖作爲資訊視覺化方法的應用技術。因此，首先以術語做爲訓練資料，進行自組織映射圖的訓練，使得圖形可以表現出領域的知識結構。訓練出自組織映射圖之後，即可根據不同的應用，將術語映射到圖形上，作爲使用者利用索引典資訊的介面。以下列出三種可能的應用：(1)瀏覽整體領域的知識結構；(2)檢索特定主題的相關術語;(3)查詢特定術語所涵攝的相關術語。這些應用的實作方法將在第 4 節中詳細說明，實際應用的例子則可參考第 5 節。

參、索引典建置

在第 2 節中我們將索引典建置的處理程序區分爲「術語選取」和「術語關係偵測」兩個部分。本節將分別就這兩個部分的相關研究以及本研究中所使用的方法提出說明。

一、術語選取

對於索引典建置，爲了提供資訊檢索與知識領域探勘等應用，所選取出來的術語需要是能具有意義且完整的語言單位。在語言學上，詞或詞組等語言單元才能夠代表特定的語意概念，所以本研究所謂的術語是指領域中代表某一概念的詞或詞組，從論文中選取出來的中英文術語必需是詞或詞組等意義完整的語言單元。但是中文在書寫上詞之間缺乏明顯的分界，辨識論文文字資訊中出現的詞與詞組相當困難；即便是英文，在詞組的辨識上也同樣困難。因此，在本研究中首先必需自論文的文字資訊中選取出同時具有單元完整且領域相關的術語。

本研究使用(林頌堅, 2002)所提出之術語選取方法。這個方法的主要概念是利用統計訊息和經驗法則，對文件資料中所有可能的字串進行篩選，過濾去可能性不高的字串，保留語言單元完整並且符合領域主題的候選術語。由於這個方法以統計訊息爲主，稍作修改後可以同時應用於多種語言的文件。進一步來說，這個術語抽取方法利用字串的前後接字複雜度來測試單元的完整性。字串前後接字的情形愈複雜，則這個字串愈可能是一個完整的術語，需要被選取出來；反之，如果這個字串不是一個完整的術語，它的前後接字複雜度必然較小。再配合上停用詞(Stop words)不能出現在術語首尾的經驗法則，可以從文件資料中抽取出多數完整的術語。另外，候選術語的主題相關性則是利用字串在所有文件出現的總次數、在出現文件中的平均出現次數和標準差等進行評估。若是候選術語出現的總次數愈高，這個術語愈有可能被用來表示領域中的重要概念，而候選術語在出現文件中的平均出現次數和標準差則用來評估這個術語在各出現文件中的主題相關性，候選術語的平均出現次數和標準差之總和愈大，表示這個候選術語愈可能與主題相關。依據上述的方法，本研究計算出現在論文文字資訊中的所有中英文字串，計算各個字串的前後接字複雜度、出現總次數、平均出現次數和標準差等資訊，並且依據這些資訊選取術語以建立領域特定的索引典。

另外，能夠進入索引典的術語應具有較好的文件鑑別性(Discrimination)，在檢索上能夠增加準確性，鑑別值(Discrimination value)便是用來衡量術語的鑑別性大小(Salton, Yang, & Yu, 1975)。對於某一個術語，其鑑別值的估算方式是以去除這個術語對資料庫文件彼此相似性的影響程度來計算。換言之，也就是計算文件相似性在術語去除前後之差。如果去除這個術語後，使得出現這個術語的文件與其他文件的相似度增加，這個術語具有較佳的文件鑑別性。事實上，文件中出現的術語，其鑑別值大多數相當接近 0，換言之從索引典中去除這些術語與否對於

檢索的效能沒有影響，所以在計算資源的考慮下，這些術語可以去除。估算出所有術語的鑑別值後，便可以找出適合的術語以建立索引典。

然而以鑑別值作爲選取術語的方法需要相當大的計算量，因此有研究者提出利用術語的文件頻率(Document frequency)，也就是術語在資料庫出現的文件數目，作爲取代鑑別值的資訊(Salton, Yang, & Yu, 1975)。資料庫中的術語可以依據它們的文件頻率分成三類：第一類是出現在相當多文件的術語，以這類術語作爲索引，將會檢索出相當多文件，對於文件缺乏鑑別性；第二類的術語僅出現在極少數的文件，但較爲罕用，因此出現的可能性不大，其主題對整個領域的代表性不佳；第三類的術語出現的文件數目在前兩類之間，同時具備適當的文件鑑別性與主題相關性，能夠符合索引的需求。依據上述的說明，在計算資源的考量下，本研究將使用文件頻率作爲術語選取的資訊。

二、術語關係偵測

在選取出術語之後，索引典建置程序將利用論文文字資訊作爲資料來源，來偵測出這些術語之間的概念關係。相關研究所提出來的術語偵測方法可以根據它們所使用的資訊分爲詞彙訊息、語法結構和統計訊息三種方法。以詞彙訊息來偵測術語關係的方法是利用某些在文件中常見的關鍵詞組來找出術語之間的概念關係(Hearst, 1998)，比方說，關鍵詞組“such as”後的術語通常是這個詞組前面術語的特定概念，關鍵詞組“and other”後的術語則通常是這個詞組前面術語的廣泛概念。以語法結構爲基礎的研究則以名詞詞組(Noun phrases)或動詞詞組(Verb phrases)爲分析對象，剖析詞組的中心語(Head)和修飾語(Modifier)，以中心語做爲詞組的一種特定概念(Grefenstette, 1997)，比方說，名詞詞組“university libraries”裡，“libraries”是中心語而“university”是修飾語，所以“university libraries”是一種“libraries”的特定概念。然而這兩種方法均依賴文字資訊中存在的術語關係來進行偵測，因此即便只出現一次的術語關係，也可以利用上述的方法偵測出來，優點是可以找出文字資訊中所有曾經出現過的術語關係，但缺點則是未能妥善利用術語關係的出現次數，區別個別術語關係的重要性，也無法剔除較不可靠的資訊。

以統計訊息爲基礎的方法則是目前在自動化建構索引的研究中最爲普遍的方法(Crouch & Yang, 1992; Park, Han, & Choi, 1995; Sanserson & Croft, 1999; Tseng, 2002)，這類方法利用術語在文件中出現的次數(Occurrences)做爲統計的訊息。當某兩個術語在文件中一起出現時，稱爲這兩個術語在此文件中具有共現(Co-occurrence)關係。如果兩個術語在文件中共現的情形愈頻繁，表示這兩個術語之間可能有愈接近的概念關係。比方說，在圖書資訊學領域中，「線上公用目錄」所出現的論文也經常出現「檢索」，因此這兩個術語具有共現關係，而且這兩個術語的概念可能相關。必須說明的是，以統計訊息估算出來的術語關係，並非是術語之間的「廣泛—特定」概念關係，而是一種在文件中術語的共現關係所顯示的樣式(Patterns)。雖然利用術語統計訊息自動化產生的索引典不比人工製成的索引典具有豐富的語意資訊，但在資訊檢索的應用上，可以利用這類索引典中的詞彙訊息做爲「問句擴展」(Query expansion)所需的資訊(Mandala, Tokunaga, & Tanaka, 1999)，提高檢索的回收率。此外，這種方法利用出現次數作爲術語關係偵測的資訊，計算迅速且不需額外的資源，因此還可以針對資訊檢索的結果動態產生的重要術語概念關係，便於使用者瀏覽與檢索相關文件資訊。

假設有兩個術語 t_a 和 t_b，並假設這兩個術語

的術語關係的估算值爲 $s(t_a, t_b)$和 $s(t_b, t_a)$，當術語關係的估算值 $s(t_a, t_b)$和 $s(t_b, t_a)$超過某一個預設的閾值時，便可推論術語 t_a和 t_b相關。利用術語的出現訊息進行術語概念關係偵測的方法可以分爲對稱式和非對稱式兩類。對稱式方法的計算結果，對於每一對術語，其間的關係是對等的，換句話說，$s(t_a, t_b)$和 $s(t_b, t_a)$的值相同；但非對稱式所偵測的術語關係不必然是對等的，也就是說 $s(t_a, t_b)$和 $s(t_b, t_a)$的估算值不一定相同。對稱式方法中最爲著名的研究是利用向量空間模式(Vector space model)的方式(Salton, 1989)，以術語在文件中出現的次數作爲特徵的基礎，產生術語的特徵向量，並計算每一對術語特徵向量的餘弦值(Cosine value)作爲術語關係的估算值，餘弦值較大的術語彼此間具有概念關係，可以作爲進一步的應用。

非對稱的方法則可以計算兩個術語之間彼此涵攝的情形(Sanderson & Croft, 1999)。給定兩個術語 t_a和 t_b，如果要計算 t_a對 t_b的涵攝關係，可以用所有出現術語 t_b的文件中同時出現 t_a的相對頻率來估算，如果相對頻率超過某一個是先給定的閾值(Threshold)，便可假設 t_a對 t_b具有涵攝關係。數學式的表示如式(1)所示，

$$s(t_a, t_b) \overset{def}{=} \frac{d_{ab}}{d_b} \qquad (1)$$

在式(1)中，$s(t_a, t_b)$表示 t_a對 t_b的涵攝關係估算值，而 d_b代表術語 t_b的文件頻率，換言之，即是 t_b出現的文件數目，d_{ab}則代表兩個術語 t_a和 t_b共同出現的文件數目。很明顯的，d_b 大於等於 d_{ab}，因此式(1)的計算結果 $s(t_a, t_b)$的值介於 0 與 1 之間。當 $s(t_a, t_b)$具有較小值的時候，術語 t_b的出現與 t_a無關。當 $s(t_a, t_b)$接近於 1 的時候，只要出現術語 t_b，必然同時會出現 t_a，此時我們可以定義 t_a對 t_b具有涵攝的關係。而且 $s(t_a, t_b)$和 $s(t_b, t_a)$不一定相等，所以涵攝關係是不對稱的，t_a 對 t_b具有涵攝關係，未必使得 t_a對 t_b具有涵攝關係。事實上，索引典裡，上層概念的術語往往對於下層的相關術語具有涵攝關係，也就是說下層的術語出現的文件中經常會有上層術語出現。以圖書與資訊學領域作爲說明，術語「網路」是「網站」的上層概念。在這個領域的相關論文中出現「網站」的論文往往也會有「網路」出現，根據式(1)，計算出「網路」對於「網站」的涵攝關係估算值，這個估算值必然接近於 1，所以「網路」對於「網站」具有涵攝關係。因此可以用涵攝關係來作爲術語關係的偵測方法。

由於涵攝關係的計算相當容易，並且計算出來的結果中包含許多術語的概念關係，因此本研究採用術語的涵攝關係作爲術語關係偵測的計算方式。

肆、索引典資訊視覺化

資訊視覺化是將資料庫中的大量資料，依據它們的資料關係產生圖形，而這個圖形可以表現出資料的特性(Card, Mackinlay & Shneiderman, 1999)。將索引典進行資訊視覺化，即是透過所產生的圖形來表示索引典中的資訊。如前面所言，索引典中蘊含的資訊不但包含術語的詞彙資訊，同時還有術語之間的概念關係。因此，本研究所採用的資訊視覺化方法需要能夠將索引典內所有的術語呈現在圖形上，而且同一圖形也能夠表現出術語之間的關係。依據可能的索引典使用情形，可以將術語之間的關係分成(1)兩個術語之間的關係(2)一組相關術語之間和(3)索引典中所有術語之間的關係等三種。對於索引典中的任何兩個術語，我們希望這兩個術語經過資訊視覺化的處理後，映射在圖形上的位置，其距離可以表現出術語之間的概念關係，術語之間愈相關，

其圖形上的距離愈近。所以當使用者需要比較索引典中的某一個術語與其他兩個術語之間的概念遠近時，透過資訊視覺化產生的圖形，比較映射結果的距離遠近，便可以了解這個術語與其他兩個術語之間的概念關係。並且依據愈相近術語其映射結果愈接近的想法，如果索引典中的一組術語，彼此都有相關的概念關係，當它們映射在圖形上時，彼此間的距離也都將會很接近，並且形成叢集(Cluster)，表現相關領域的某一個主題。所以當使用者想利用這個圖形介面檢索索引典中某一術語的所有相關術語時，便可以檢視圖形上映射在這個術語附近的所有術語，即能符合他的需求。當索引典上所有的術語依據它們之間的關係映射到圖形時，各組主題相關的術語會在圖形上形成各個叢集，叢集所代表的主題便是這個領域中的重要主題，而各個叢集之間的距離與關係則形成了領域各個主題之間的結構，因此所產生的圖形便可以表示領域的所有主題與整體知識結構。

過去的研究中，對於術語或是文件等文字資料進行資訊視覺化處理，通常的做法是將每一筆文字資料表示成一組特徵向量(林頌堅, 2004a)。接著利用 SVD(Singular Value Decomposition) (Landauer, Laham, & Derr, 2004)、PCA (Principal Component Analysis)或是 MDS (Multidimensional Scaling) (Huang, Ward, & Rundensteiner, 2003)等統計導向的方法或是自組織映射圖的類神經網路(Artificial neural network)導向方法(Flexer, 2001)將這些特徵向量映射到圖形上。統計導向方法需要及大量的運算資源並且新增資料無法相容於先前產生的結果，在實作方面較不方便。因此，本研究採用自組織映射圖作爲索引典資訊視覺化的技術。經過充分的訓練之後，自組織映射圖技術所產生的圖形可以將高維度資料項映射在二維圖形上，並且盡量保持資料項之間的關係，將較相似的資料項映射到距離較近的位置，使得相關的資料項形成叢集，進而呈現出整個領域的主題與結構。在過去，作者曾利用自組織映射圖技術對領域主題探勘進行了一序列的研究(林頌堅, 2004a, 2004b)，來發掘知識領域的重要主題以及它們的發展趨勢，在本研究中則應用這項技術進行索引典資訊視覺化。

一、自組織映射圖訓練

自組織映射圖的運作概念是利用一組排列成方陣的節點(Nodes)來表示輸入的資料項和它們之間的關係(Kohonen, 1989)。每一個資料項和圖形上的每一個節點都以一組特徵向量來代表。在還沒經過訓練前，節點特徵向量上的特徵值是隨機指定的，所以整個自組織映射圖是沒有組織的狀態。輸入資料之後，開始進行重複多次的訓練過程。每次隨機選擇一個資料項(Data item)，以它的特徵向量與圖形上所有節點的特徵向量進行比對，計算兩者間的歐幾里德距離。根據比對的結果，選擇與資料項特徵向量距離最小的節點與在這個節點鄰近範圍內的節點進行調適(Adaptation)，縮小這些節點的特徵向量與資料項特徵向量的距離。使節點特徵向量相似於資料項特徵向量，並且使得鄰近範圍內節點的特徵向量也彼此相似。經過多次訓練後，自組織映射圖上的節點將會逐漸組織起來，使得特徵向量接近的資料項映射到同一節點或鄰近的節點上，自組織映射圖便可以表現出資料項之間的關係。借助這樣的特性，自組織映射圖可以將高維度特徵向量的資料項映射到圖形上，做爲資訊視覺化的工具。

本研究將以索引典中的術語作爲自組織映射圖的訓練資料，並且以術語和各術語的共現程度作爲特徵向量中的特徵值。術語之間的共現程度則是以術語在各論文中的出現次數爲基礎，利

用向量空間模式的餘弦值來計算(Salton & McGill, 1983)，如式(2)所示。

$$c(t_a,t_b) \overset{def}{=} \frac{\sum_{i=1}^{N}(f_{ai} f_{bi})}{\sqrt{\sum_{i=1}^{N} f_{ai}^{\,2}}\sqrt{\sum_{i=1}^{N} f_{bi}^{\,2}}} \qquad (2)$$

式(2)中，$c(t_a, t_b)$表示術語 t_a 和 t_b 在所有文件中的共現程度，則 f_{ai} 和 f_{bi} 分別代表 t_a 和 t_b 在第 i 筆文件中的出現次數，而 N 則是代表文件的數目。以式(2)來說，兩個術語愈常一起出現在文件中，其共現程度愈大；若反之，則愈小。接下來，術語 t_a 的特徵向量 F_a 定義如式(3)。

$$F_a = \begin{bmatrix} c(t_a,t_1) \\ \Lambda \\ c(t_a,t_j) \\ \Lambda \\ c(t_a,t_M) \end{bmatrix} \qquad (3)$$

在式(3)中，$c(t_a, t_j)$是術語 t_a 與第 j 個術語 t_j 的共現程度，並假定共有 M 個術語，特徵向量 F_a 便是 t_a 與各術語共現程度的分布情形。若是某一個術語與各術語的共現程度分布情形和另一個術語相似，表示這兩個術語與某一群術語常出現在同一文件中，這兩個術語與這群術語間都相關，換句話說，這兩個術語之間也很有可能相關；而且因爲兩個術語具有相似的共現程度分布情形，它們特徵向量之間具有較小的歐幾里德距離。相反的，若是共現程度分布情形不相似，則特徵向量之間的距離將會較大，這兩個術語即是不相關。所以可以利用術語與各術語之間的共現程度來定義特徵向量，而使得這個特徵向量適合應用於自組織映射圖的訓練。但因爲在許多文件中相關術語間不一定有共同出現的情形，因此利用LSA(Latent Semantics Analysis)技術(Deerwester et. al., 1990)對相關程度進行平滑化(smoothing)。

在選擇訓練資料時，最爲簡便的方式是以全部的術語作爲訓練資料，依據標準的自組織映射圖訓練方式，每次隨機選擇一組特徵向量進行調適，並且逐步縮小調適的幅度與範圍。由上述訓練過程的說明，我們可以知道訓練資料項的選取順序將會影響結果的成效與效果。在索引典資訊視覺化的應用裡，術語特徵向量的特徵值分布情形差異相當大，在概念階層下層的術語概念較特定，並且只與極少數的術語有共現關係，特徵向量相當獨特；但在概念階層上層的術語，則與許多術語都具有共現關係，這類術語的特徵向量彼此相似，很難藉由特徵向量來區別術語的不同。因此，以全部術語做爲訓練資料的方式所得到的結果並不穩定，必需審慎選擇訓練資料。如果只選擇上層概念的術語，則由於特徵向量彼此間差異不大，產生的自組織映射圖無法根據資料的特性，將所有的術語映射到適當的位置。反之，只選擇下層概念的術語則特徵向量差異過大，自組織映射圖無法獲得充分的訓練，不足以代表領域全體的主題與知識結構。因此要能代表索引典內多數術語，所選取的術語以索引典概念階層的中間層級術語較合適。在本論文中，我們將被選取出來參與訓練的術語稱爲核心術語。

前面的「索引典建置」程序中利用術語之間的涵攝關係來偵測術語關係，決定索引典的概念階層，因此這個處理程序便利用術語間的涵攝關係來選取核心術語。以涵攝關係來看，在概念階層上層的術語涵攝其他術語的情形較多；反之，在概念階層下層的術語則被其他術語涵攝的情形較多。因此，在選取核心術語時，便可以根據術語彼此間的涵攝情形，選擇在概念階層中間層

級的術語作爲核心術語。選取出核心術語之後，便以這些核心術語進行自組織映射圖訓練。

二、術語映射

索引典視覺化流程的最後一個步驟是將索引典中所有的術語映射到自組織映射圖上。如式(2)或式(3)的方式定義所有術語的特徵向量，並計算術語特徵向量與訓練好的自組織映射圖上每一個節點特徵向量之間的歐幾里德距離，選擇距離最小的節點做爲術語映射的節點。此時，索引點上所有的術語便會依據它們間的距離映射到相對應的節點上。

在這裡我們舉出三種應用情形，討論各種情形下術語的映射方式。

(一)瀏覽整體領域的知識結構。當使用者想要瀏覽整個領域的知識結構時，可以將領域中的核心術語映射到圖形上，呈現出整體的概念分布情形。

(二)檢索特定主題的相關術語。當使用者想要深入局部的知識結構或利用索引典進行資訊檢索的應用時，便可以先根據整體的自組織映射圖，根據核心術語所形成的主題叢集，選定一個範圍，將這個範圍內的所有術語映射到圖形上。

(三)查詢特定術語所涵攝的相關術語。如果使用者對某個術語所代表的概念感到興趣，可以輸入這個術語，查詢這個術語所涵攝的術語，將這些術語分別映射到圖形上。此時使用者便可以依據這些術語在自組織映射圖上映射的位置以及和查詢術語的距離，推斷查詢術語的概念。

伍、實驗與結果

爲了驗證本論文所提出方法的成效，將這個方法實際應用於《圖書與資訊學刊》中發表的論文。《圖書與資訊學刊》的內容以圖書館學、目錄版本學、資訊科學、檔案學、博物館學等相關論著爲主，迄今(2005 年)已經出版了 52 期，是國內圖書資訊學中長期出版並且有代表性的刊物，以該期刊裡的論文做爲研究對象，將可以了解國內圖書資訊學領域的重要術語、相關主題以及術語之間的關係。在過去，作者已經針對這份期刊進行過術語抽取(林頌堅, 2002)以及主題分析(林頌堅, 2003)等研究，本論文將以前面的研究爲基礎，進一步探討國內圖書資訊學的知識結構。

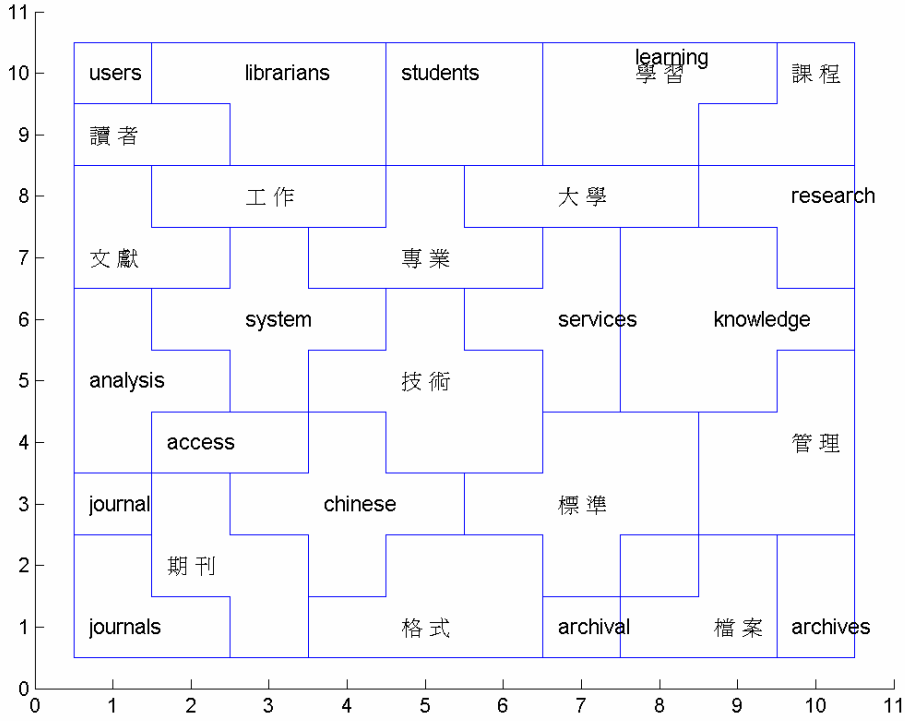
本研究蒐集了《圖書與資訊學刊》第 16 期到第 52 期的論文資料，共 256 篇。將論文的發表時間、作者、中英文題名和摘要等資料建置成資料庫，再針對題名和摘要等文字資料，抽取關鍵術語，並且計數這些術語的統計訊息。在本研究術語抽取的過程中，將術語的出現總次數設定爲 15 次以上，術語在出現論文中的平均出現次數與標準差的和必須在 3.0 以上，左右接字的複雜度則設定爲 1.0 以上，而術語的最小文件頻率則設爲 3，結果共選擇出 209 個術語。在術語關係的偵測上，本研究以術語 ta 對於另一術語 tb 的涵攝關係估算值大於等於 0.5 以上做爲 ta 涵攝 tb 的情形，並且計算各個術語涵攝其他術語的數目。其中，涵攝較多術語的包括「圖書館」(涵攝 165 個術語)、“library”(141 個)、“information”(111 個)、「研究」(99 個)和「資訊」 (82 個)等等，很明顯這些術語在《圖書與資訊學刊》中是屬於最上層概念的術語。

接下來進行自組織映射圖的訓練。首先以術語的涵攝情形選擇合適的核心術語，本研究將涵攝術語的數目在 4 到 10 之間的術語做爲是核心術語，核心術語的總數共有 28 個，建立這 28 個核心術語的特徵向量。本研究使用的自組織映射圖的規模爲 10×10，利用核心術語對自組織映射圖的，訓練次數爲 1000 次。訓練出自組織映射

圖後，便可以利用此一組織映射圖做爲資訊檢索及領域知識探勘的介面。以下以實例說明前述所提出的三種可能的應用情形。

一、瀏覽整體領域的知識結構

當使用者想瀏覽《圖書與資訊學刊》中重要的研究主題以及主題之間的關係時，可以將核心術語映射到訓練好的自組織映射圖，透過術語在圖形上的映射結果了解領域的知識結構。圖二中呈現的圖形便是以核心術語映射到自組織映射圖上所產生的圖形。



圖二：本研究將《圖書與資訊學刊》的核心術語映射到自組織映射圖上所產生的圖形

在圖二上，術語呈現在映射的節點上，比方說，術語“journals”映射的節點爲(1,1)，「期刊」則映射在(2,2)的節點上。從“journals”、“journals”和「期刊」等互爲單複數型或翻譯的術語分別映射在鄰近節點(1,1) 、(1,3)和(2,2)上，可見得本研究所提出的索引典資訊視覺化方法，能夠將相關的術語映射到圖形鄰近的位置上，因此可以表現出領域的知識結構。同樣的情形還有被映射到(7,1) 、(9,1)和(10,1)等節點上的“archival”、「檔案」和“archives”等相關術語。另外，本研究並將特徵向量相接近的節點被群組起來，使得術語映射的結果更加清楚，比方說，圖二的節點(1,1)、(1,2)和(2,1)等被群組起來，並且由節點(1,1)上映射的術語“journals”，可以知道這個群組裡的節

點都與期刊的概念相關。

透過圖二的觀察，我們可以發現許多《圖書與資訊學刊》常出現的主題，包括期刊、文獻、讀者、使用者(Users)、系統(System)、圖書館館員(librarians)、格式、檔案、知識(Knowledge)、學習、課程等等。更進一步地，我們可以發現許多圖上表現出的術語相關性。這些術語之間的關係，原本以人工的方式很難確認出來，然而藉由術語關係偵測與索引典資訊視覺化等程序處理後，讀者可以很容易地利用自組織映射圖，發現這些關係。比方說，與「檔案」此一概念相關的核心術語，共有三個：“archival”、「檔案」和“archives”。如前所述，這三個術語在自組織映射圖上都被映射到右下方的節點，而且映射在這些節點附近的核心術語包括了「管理」、「標準」和「格式」。在檢索《圖書與資訊學刊》的相關論文後，我們可以發現目前檔案相關的研究著重於檔案描述以及檔案管理兩個部分。因此，本研究的結果正符合相關論文的主題。

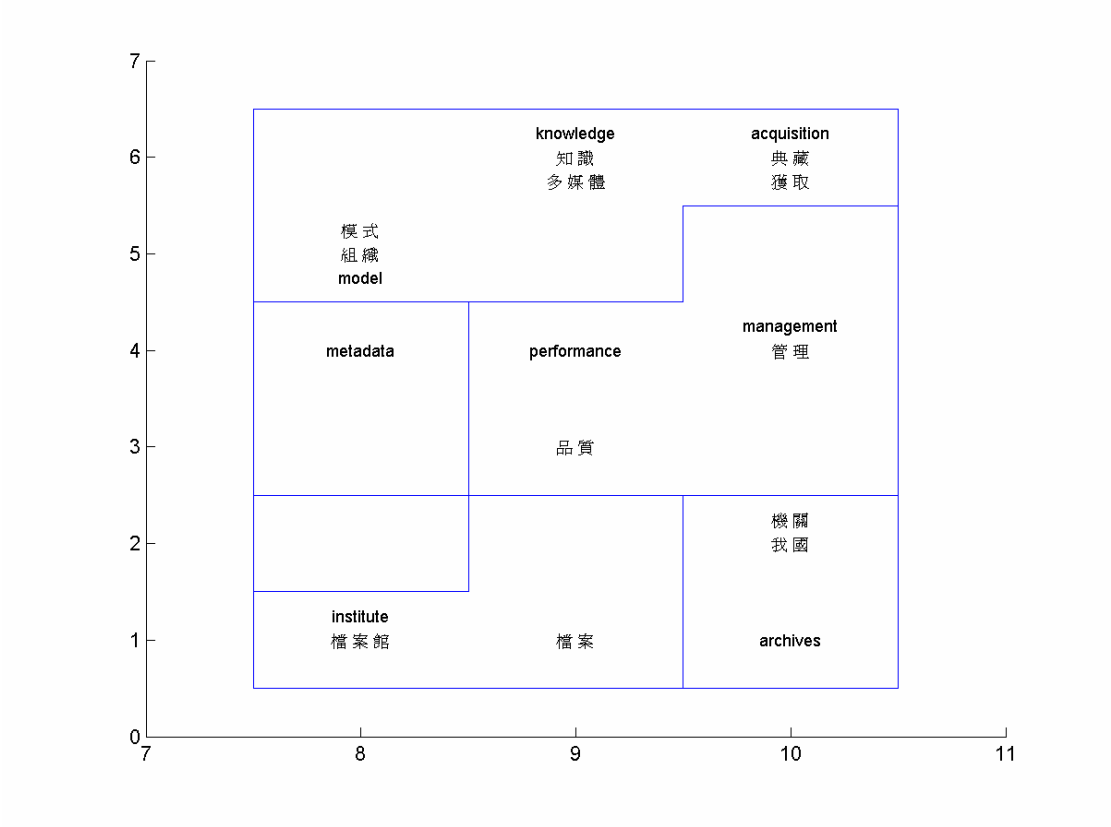
此外，圖書資訊學課程的規劃與設計以及圖書館館員的專業與工作也是圖書資訊學領域中相當重要的研究課題。圖二中我們可以觀察到課程規劃與設計相關的術語映射在圖形的右上方，包括節點(8,10)上的「學習」和“learning”以及(10,10) 上的「課程」等術語。圖形上與這些術語接近的術語則有“students”、「大學」和“research”等等，並且很清楚地這些術語彼此間相關。圖書館館員的專業與工作等術語則是映射在圖形左上方的節點上，包括節點(3,10)上的“librarians”、節點(3,8)上的「工作」和節點(5,7)上的「專業」等等，相關的術語則有“users”、「讀者」、“system”、「技術」和“services”等等。

從上述的例子可以發現經過資訊視覺化的處理之後，領域中的重要主題可以清楚地從圖形上觀察得到；並且透過直覺的圖形顯示，使用者也可以發現許多原本不容易理解的概念關係。因此，可以利用這樣的結果做爲瀏覽整體領域知識結構的方法。

二、檢索特定主題的相關術語

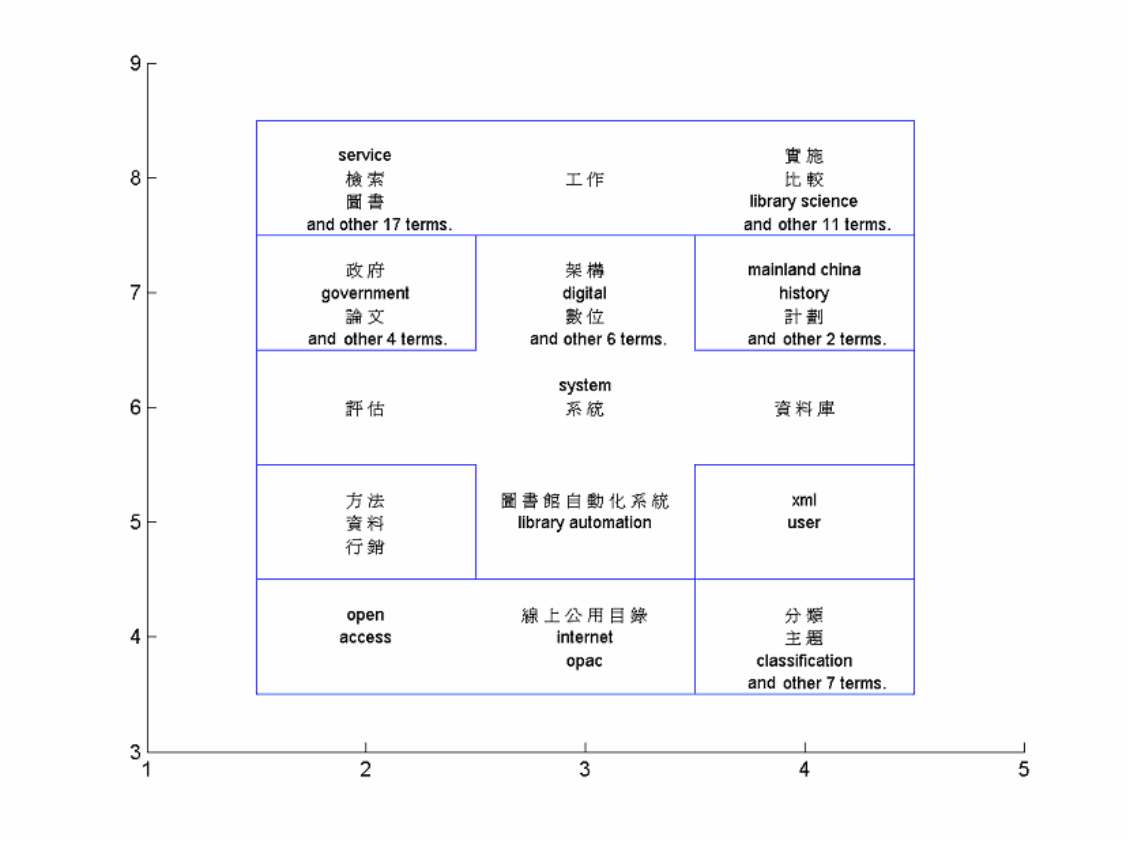
本研究將應用索引典建置及索引典資訊視覺化處理後所產生的結果，檢索《圖書與資訊學刊》中與管理和技術主題相關的術語。以下便是檢索的過程與結果。

管理的意義是對各種資源進行評估、安排以及有效的控制，以提高生產的品質並且增加績效，因此近來這個議題被資源日益緊縮的圖書館界所重視，也因此圖書資訊學領域有相當多的論文探討這個議題。在圖二上，術語「管理」映射的位置是在右下的節點(10,4)，因此鄰近範圍內的術語都將與這個主題相關。爲了更進一步了解相關的概念以及「管理」和其他術語之間的關係，我們以節點(8,1)到(10,6)爲範圍，檢索這個範圍內的所有術語。檢索的結果如圖三所示。



圖三：本研究節點(8,1)到(10,6)範圍內所有術語的檢索結果

在圖三中，除了映射在節點(10,4)上的術語「管理」和“management”之外，最靠近的術語爲映射在節點(9,4)上的術語“performance”和節點(10,4) 上的「品質」。很明顯的，這兩個術語與管理的概念十分相關，在《圖書與資訊學刊》中探討管理主題的相關論文，有些也會提到成效和品質的相關概念。《圖書與資訊學刊》中管理主題的相關論文還包括了檔案管理和知識管理兩方面。在檔案管理方面，由於近來檔案研究的成長、對於檔案保存及利用的重視以及檔案局的成立，在《圖書與資訊學刊》中有相當數量的相關論文發表。在圖三下方的(8,1)、(9,1) 、(10,1) 和(10,2)等節點上，我們可以觀察到相關的術語包含了“institute”、「檔案館」、「檔案」、 “archives”、「機關」和「我國」等等。另一方面，從圖三可以觀察到，管理與知識相關的術語包含了“acquisition” 、「典藏」和「獲取」等，這些術語表現了《圖書與資訊學刊》論文在知識管理研究中，主要著重在知識的典藏與獲取上。所以應用上述的方法可以產生「管理」相關的術語，對於檢索管理概念相關論文必然有所助益。



圖四：本研究節點(2,4)到(4,8)範圍內所有節點的檢索結果

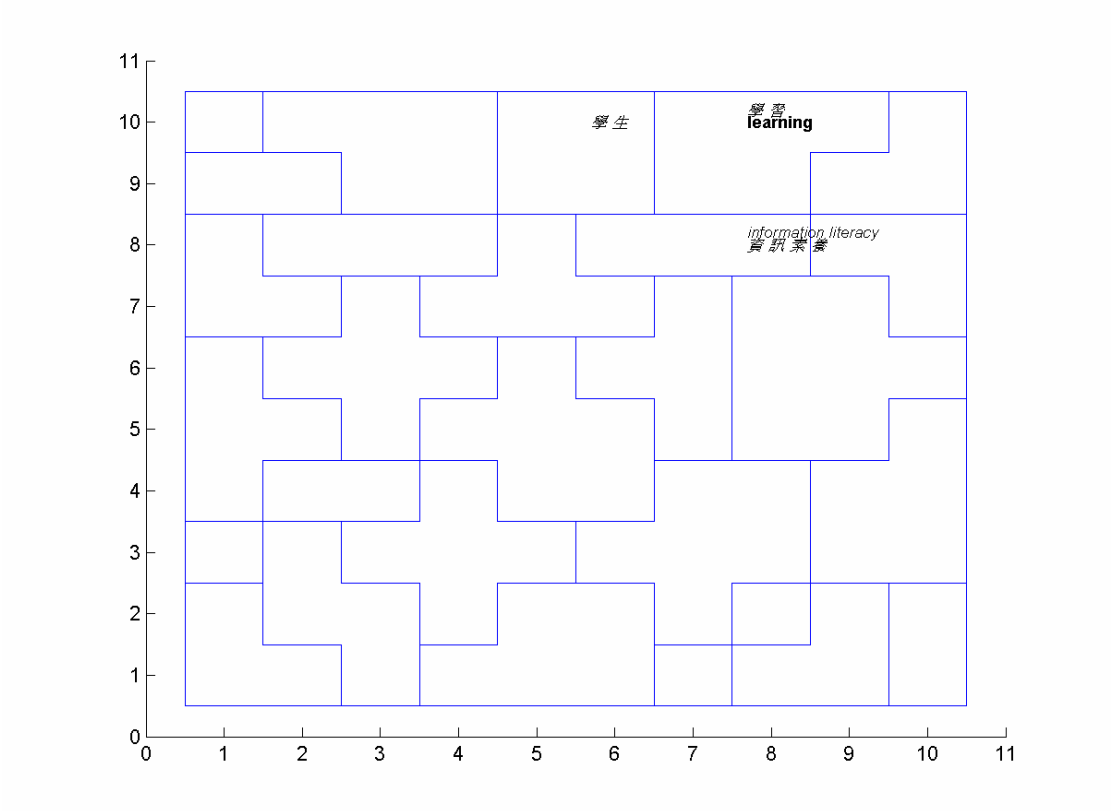
近年來，由於資訊科技的快速發展，利用電腦與網路系統來儲存與傳遞資料，已經是每一個圖書館與每一位圖書資訊專業人員所必須面對的課題，圖書資訊學領域中有極爲大量的論文在探討這一方面的主題，提出各種現在與未來的資訊處理系統的建議方案、導入與成效評估。因此，接下來我們檢索術語「系統」映射節點附近的術語，來觀察圖書資訊學領域中有關系統的術語。圖二上，術語「系統」映射的節點爲(3,6)，因此我們檢索節點(2,4)到(4,8)範圍內的所有術語，結果如圖四。在圖四中由於某些節點檢索出來的術語數量較多，爲了避免增加使用者的認知負擔，容易檢視，我們僅列出出現總次數較多的前三個術語。比方說，節點(2,8)中僅列出“service”、「檢索」和「圖書」等三個出現總次數最多的術語，其餘出現總次數較少的 17 個術語不加以列出。

從圖四上的術語映射結果，可以觀察到《圖書與資訊學刊》論文與系統相關的術語包括了「檢索」、「論文」、「評估」、「數位」、「圖書館自動化系統」、「線上公用目錄」、“xml”和「分類」等等，這些都是圖書資訊學領域中常與「系統」相提並論的術語。在檢索《圖書與資訊學刊》的論文時，使用者可以利用這些術語提升檢索的效能。

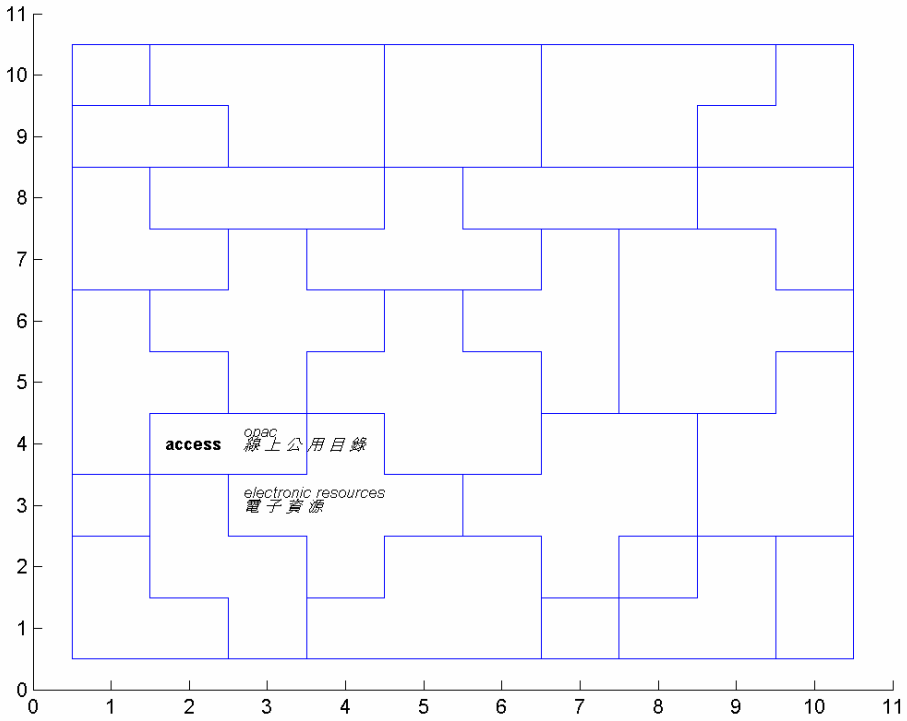
三、查詢特定術語所涵攝的相關術語

接下來將利用特定的術語查詢它所涵攝的關術語。本研究以“learning”和“access”兩個術語爲例，查詢這兩個術語所涵攝的相關術語。圖五和圖六分別是它們的查詢結果，在圖上，黑體字所表示的是查詢術語的名稱及映射位置，在本研究的例子裡分別是“learning”和“access”，而斜體字則表示查詢術語所涵攝的術語名稱及映射位置。

在圖五上，可以觀察到“learning”的涵攝術語，除了自己本身的中譯「學習」以外，還包括「資訊素養」、“information literacy”和「學生」等術語，並且依據節點群組可以將這些術語分爲兩組，一組爲互爲翻譯的「資訊素養」和“information literacy”，另一組則爲「學生」。圖六則表示“access”的涵攝術語也可以分爲兩組互爲翻譯的術語，“opac”和「線上公用目錄」，以及“electronic resources”和「電子資源」。當使用者檢索《圖書與資訊學刊》的論文時，可以利用上述的方法，縮小檢索的主題範圍。以“access”的例子來說，當使用者以這個術語做爲查詢問句，若是檢索結果的資料太多，使用者便可透過上述的方法，判斷他的需求與線上公用目錄或電子資源較相關，使用相關的涵攝術語來進行檢索。



圖五：本研究中“learning”之涵攝術語的查詢結果



圖六：本研究中“access”之涵攝術語的查詢結果

陸、結論

由於科學研究的急遽成長，產生許多優異的研究成果，不僅造成原有學術領域的知識結構快速地變動，並且產生許多新的學術領域，對學術研究者來說，提供包含領域重要術語以及概念關係的索引典，可以增進相關文獻檢索的效率，也能夠幫助他們認識領域的研究問題、方法、技術和理論等知識結構。另一方面，資訊科技的進步，使得索引典中的詞彙資訊可以透過電子形式呈現，提供更爲直覺而有效率的檢索方式，使得使用者能夠快速而便利地取得他們需求的資訊。針對上述的問題，本論文提出一系列索引典自動化建制與資訊視覺化方法，利用相關論文的文字資料作爲資訊來源，以術語出現於文字資料的統計訊息爲基礎，選取重要的術語以及偵測術語之間的概念關係，並且以自組織映射圖技術做爲資訊視覺化的方法，將索引典中所儲存的詞彙資訊表示成具有意義的圖形。這種術語排列方式相較於傳統循序的線狀排列或階層式的樹狀排列不但可以表示更多的資訊，而且在使用上更直覺化。可以將術語的主題關係表現在圖形上，達到資訊視覺化的效果。由於這些方法都是由資料驅動(Data driven)，結果可以隨輸入資料自動調適，容易擴充，因此所需計算資源也不會太大，適合用於現代科技知識與資訊發展相當快速的學術領域。

本論文並且以政治大學圖書館所出版的《圖

書與資訊學刊》爲例，利用學刊論文的題名與摘要等文字資料，建置索引典，再利用這些術語與概念關係，進行資訊視覺化。在本論文中不但產生了索引典的二維圖形表示，並且以實例說明了這個結果在(1)瀏覽整體領域的知識結構、(2)檢索特定主題的相關術語、(3)查詢特定術語所涵攝的相關術語等等在資訊檢索以及領域知識探勘方面的應用。

（收稿日期：2005 年 8 月 4 日）

參考書目：

Card, S. K., Mackinlay, J. D. & Shneiderman, B. (1999). Information visualization. Readings in information visualization—Using vision to think, 1-34. Morgan Kaufmann.

Crouch, C. J. & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. Proceedings of the 15nd ACM SIGIR Conference on Research and Development in Information Retrieval, p.77-88.

Deerwester, S., et. al. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. Intelligent Data Analysis, 5(5), 373-184.

Greffenstette, G. (1997). SQLET: Short query linguistic expansion techniques, palliating one-word queries by providing intermediate structure to text. Proceedings of RIAO, 500-509.

Huang, S. Ward, M. O. & Rundensteiner, E. A. (2003). Exploration of dimensionality reduction for text visualization. Technical Report TR-03-14, Worcester Polytechnic Institute, Computer Science Department.

Hearst, M. A. (1998). Automated discovery of WordNet relations. In Cbristiane Fellbum (Ed), WordNet: an electronic lexical database. Cambridge, MA: MIT Press.

Kageura, K. & Umino, B. (1996). Methods of automatic term recognition—A review. Terminology, 3(2), 259-289.

Kohonen, T. (1989). Self-organization and associative memory. New York: Springer-Verlag.

Landauer, T. K., Laham, D. & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. Proceedings of the National Academy of Science of the USA, 101, 5214-5219.

Mandala, R., Tokunaga, T., & Tanaka, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, 191-197.

Merkl, D. (1997). Exploration of text collections with hierarchical feature maps. Proceedings of the 20nd ACM SIGIR Conference on Research and Development in Information Retrieval, 186-195.

Park, Y., Han, Y., & Choi K. (1995). Automatic thesaurus construction using Bayesian networks. Proceedings of CKIM’95, 212-217.

Rowley, J. E. (1992). Organizing knowledge. New York: Ashgate Publishing Limited.

Sanderson, M. & Croft, B. (1999). Deriving concept hierarchies from text. Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, 206-212.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of information by computer. Reading, MA: Addison-Wesley.

Salton, G. & McGill, M. J. (1983). Introduction to modern information Retrieval. New York: McGraw-Hill.

Salton, G., Yang, C. S. & Yu, C. T. (1975). A Theory of term importance in automatic text analysis. Journal of the American Society for Information Science, 26(1), 33-44.

Soergel, D. (1985). Organizing information—Principles of data base and retrieval systems. New York: Academic Press, Inc.

Tseng, Y-H. (2002). Automatic thesaurus generation for Chinese documents. Journal of the American Society for Information Science and Technology, 53(13), 1130-1138.

林頌堅(2002)。圖書與資訊學刊的高頻詞語抽取與分析。圖書與資訊學刊，42，15-28。

林頌堅(2002)。基於詞語抽取的圖書與資訊學刊研究主題分析。圖書與資訊學刊，47，15-35.

林頌堅(2004 a)。以自組織映射圖進行計算語言學領域視覺化之研究。Proceedings of ROCLING XVI(第十六屆自然語言與語音處理研討會論文集)，69-77。

林頌堅(2004 b)。以自組織映射圖探勘計算語言學研究發展之趨勢。2004 年現代資訊組織與檢索研討會，69-77。